Dating and Stratifying a Historical Corpus with a Bayesian Mixture Model

Hellwig, Oliver

Dating and Stratifying a Historical Corpus with a Bayesian Mixture Model

Oliver Hellwig
Department of Comparative Linguistics
University of Zurich
hellwig7@gmx.de

Abstract

This paper introduces and evaluates a Bayesian mixture model that is designed for dating texts based on the distributions of linguistic features. The model is applied to the corpus of Vedic Sanskrit the historical structure of which is still unclear in many details. The evaluation concentrates on the interaction between time, genre and linguistic features, detecting those whose distributions are clearly coupled with the historical time. The evaluation also highlights the problems that arise when quantitative results need to be reconciled with philological insights.

Keywords: Textual chronology, Bayesian mixture model, Vedic Sanskrit

1. Introduction

While the historical development of the classical Chinese and European (Latin, Greek) literature is well understood, the chronology of ancient corpora from the Near and Middle East (Sumerian, Egypt, Hebrew) as well as from South Asia is often heavily disputed. The situation is especially complicated for the Vedic corpus (VC) of ancient India. Vedic is the oldest form of Sanskrit, an Indo-Aryan language that is the predecessor of many modern Indian languages (Masica, 1991, 50–53). The VC presumably has been composed between 1300 and 400 BCE, and consists of metrical and prose texts that describe and discuss rituals and their religious significance (Gonda, 1975; Gonda, 1977). Being a large sample of an old Indo-European language, the VC often serves as a calibration point in diachronic linguistic studies. Moreover, it provides the foundations for the major religious and philosophical systems of India. Therefore, it is important to have a clear idea of its temporal axis.

Studying the diachronic linguistic development of Vedic is challenging, because external historical and archaeological evidence is unclear, missing or has not been explored so far (Rau, 1983; Witzel, 1995), and the texts do not provide datable cross-references. The situation is further complicated by the lack of reliable authorial information and of old manuscripts or even autographs (Falk, 1993, 284ff.), as well as by the fact that many, or even all, ancient Indian texts, in their current form, have been compiled from different sources or may have originated from oral literature. Moreover, even the Rigveda (ṚV), the oldest Vedic text, shows traits of an artificial language that was no longer in active use (Renou, 1957, 10). While it is easy to distinguish Old from Middle English just by reading a few lines of text, diachronic linguistic changes in post-Rigvedic Sanskrit are difficult to detect with philological methods. As a consequence, dates proposed for individual texts in the secondary literature can differ by several hundreds of years or are often not given at all.

In spite of these difficulties, 150 years of Vedic studies have produced a coarse chronology of the VC. This paper introduces a Bayesian mixture model called ToB ("time or background") that refines and clarifies this chronology. While most Bayesian mixture models with a temporal component focus on deriving linguistic trends from known temporal information (see Sec. 2.), the model proposed in this paper takes the opposite approach and derives temporal information from linguistic features. For this sake, it integrates the current state of knowledge in the text-historical domain as a subjective Dirichlet prior distribution, and models refined dates of composition with a hidden temporal variable. Non-temporal factors that may influence the linguistic form of texts are modeled with a background branch (Chemudugunta et al., 2007), and the decision between time or background is based on the subtypes of linguistic features.

This design choice is due to the philological and text-historical orientation of the model: An important aspect of its evaluation consists in finding linguistic features that can serve as diachronic markers in Vedic. Most research has concentrated on the ṚV as the oldest Vedic document and on rare linguistic features that disappear soon after the Rigvedic language (e.g., the subjunctives of all tenses). These studies are therefore of limited use for dating later Vedic texts. This paper uses a broader range of features including lexical as well as non-lexical ones, which are generally assumed to be less dependent from the topic of texts (Stamatatos, 2009; Mikros and Argiri, 2007). By inspecting the conditional distributions of the trained model, I will show that simple linguistic features such as, for instance, the frequencies of certain POS n-grams are good predictors of time, as they reflect changing syntactic preferences in late Vedic texts. The underlying syntactic developments were discussed in linguistic studies (see Sec. 2.) as well as in recent publications using quantitative frameworks (Hellwig, 2019).

Regarding the role of background distributions, the interaction between linguistic surface and non-temporal factors such as the genre (Hock, 2000; Jamison, 1991) or the place of origin of a text (Witzel, 1989) is well known, but has not been assessed in a quantitative framework in Vedic studies so far. The design of the model discussed in the paper provides a principled approach for distinguishing between time-related features and those that are generated by non-temporal factors. The latter can serve for extending future versions of the proposed model with further non-temporal hidden variables. Section 5.1. will show that the genre of

Vedic texts is a prime candidate for such an extension.

The rest of the paper is structured as follows. After a brief overview of related research in Sec. 2., Sec. 3. sketches the model and Sec. 4. describes the data used in this paper. The main part of this paper (Sec. 5.) deals with the evaluation of the results. The problem formulation itself – refining a disputed chronology of texts – implies that there is no accepted gold standard for the extrinsic evaluation of the model. Since the composition of the ṚV, the oldest and most famous Vedic text, has been studied extensively in previous research, Sec. 5.4. uses this text as a test case for a detailed philological evaluation of the model results. Section 6. summarizes this paper and discusses future extensions of the proposed model. – Data and script are available at `https://github.com/OliverHellwig/sanskrit/tree/master/papers/2020lt4hala`.

2. Previous Research

Vedic studies have examined the temporal structure of the VC for more than 150 years, starting with a chronology that is tightly coupled with the content of texts and implicitly still used in many publications (Levitt, 2003). Since external historical evidence is not available, linguistic features, the meter and the content were used as chronological markers for studying the temporal structure of the ṚV (Avery, 1872; Lanman, 1872; Arnold, 1905). Large parts of the post-Rigvedic corpus were only sporadically considered in diachronic studies. Most scholars concentrated on limited sets of words (Wüst, 1928; Poucha, 1942) or morpho-syntactic features they assumed to indicate the old or young date of a text. These features include variations in the frequencies of case terminations (Lanman, 1872; Arnold, 1897a) or verbal moods (Arnold, 1897b; Hoffmann, 1967; Kümmel, 2000). Witzel (1989) extended the set of diachronically relevant features and studied the relationship between geographical clues found in the texts and their linguistic form. More recently, a limited number of publications applied statistical (Fosse, 1997), information theoretic (Anand and Jana, 2013), and discriminative machine learning methods (Hellwig, 2019). As the temporal granularity of quantitative results is often much coarser than expected by philologists, reconciling these results with traditional scholarship remains an open problem.

Many NLP papers that deal with diachronic data do not focus on the temporal information as such, which is assumed to be known. Instead, they use it to detect, for example, semantic changes in diachronic corpora (Kim et al., 2014; Hamilton et al., 2016; Frermann and Lapata, 2016) or the historical distribution of ideas (Hall et al., 2008). Several authors have integrated temporal information into mixture models either by imposing constraints on the mixture parameters (Blei and Lafferty, 2006) or directly sampling time stamps of documents from a continuous distribution (Wang and McCallum, 2006). As it is often difficult to decide if linguistic variation inside a text is due to time or to different authors, models for authorship attribution as proposed by Rosen-Zvi et al. (2004), Seroussi et al. (2012) or, with a Dirichlet process, Gill and Swartz (2011) are equally relevant for this paper.

3. Model

Linguistic variation in historical corpora spanning a long time range can be due to diachronic changes in the language as well as to other factors such as different textual styles, genres or geographic variation. The model proposed in this paper accounts for these causes of linguistic variation by combining two admixture sub-models (see Fig. 1). The first of these sub-models, which is responsible for sampling the latent time variable $\mathbf{t}$, obtains a subjective time prior $\boldsymbol{\tau}$. The second sub-model is initialized with an uninformative prior $\boldsymbol{\alpha}$ and represents background distributions, which are meant to capture non-temporal trends in the data (Chemudugunta et al., 2007).

When a token x_{dku} of feature subtype k (e.g. case=accusative) is sampled in document d, its feature type decides if it is drawn from the time related distribution $\boldsymbol{\theta}_{t_{dku}}$ or from a background distribution $\boldsymbol{\psi}_{s_{dku}}$. This approach differs from the one proposed by Chemudugunta et al. (2007), where the sampling path is chosen on the basis of document distributions. Since this paper focusses on the diachronic distribution of features, this design decision is considered a relevant part of the model.

The latent discrete time variable $\mathbf{t}$, which denotes the true (but unknown) dates of composition of individual text sections, is split into 30 time bins. The size of these bins corresponds to slices of approximately 35 years, a value often assumed to span one generation of authors. Results of previous text-historical research (see Sec. 4.) are integrated using a section-wise subjective Dirichlet prior $\boldsymbol{\tau}_d$ of the latent time variables $\mathbf{t}$, which represents text-historical knowledge about the approximate dates of composition of each text section. For constructing this prior, text-historical information, as listed in Sec. 4., is first encoded as a range of section-wise lower and upper dates l_d, u_d. Value i of the prior τ_d (representing the prior of time bin i for text section d) is then modeled using the cumulative density function (cdf) of a Normal distribution with $\mu_d = \frac{1}{2}(l_d + u_d)$ and $\sigma_d^2 = (u_d - mu_d)/z_d$. The z-value z_d is chosen such that l_d and u_d represent the lower and upper limits of the 70% confidence interval of the corresponding Normal distribution. The prior can now be calculated as the difference of the cdfs of two adjacent time bins:

$$\tau_{di} = \text{cdf}\big(\mathcal{N}(i|\mu_d, \sigma_d^2)\big) - \text{cdf}\big(\mathcal{N}(i-1|\mu_d, \sigma_d^2)\big) \quad (1)$$

Using standard Dirichlet integration and the notation given in Fig. 2, the posterior predictive for a collapsed Gibbs sampler can be obtained from the joint distribution of all variables by integrating out the variational parameters $\boldsymbol{\Omega} = \{\boldsymbol{\omega}, \boldsymbol{\phi}, \boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{\psi}\}$ (see Fig. 1 for details):

$$\begin{aligned} &p(t_n, s_n, g_n|\mathbf{t}^{-n}, \mathbf{s}^{-n}, \mathbf{g}^{-n}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}) \\ &= p(\mathbf{t}, \mathbf{s}, \mathbf{g}|\boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}) \\ &= \int_{\boldsymbol{\Omega}} p(\mathbf{t}, \mathbf{s}, \mathbf{g}, \boldsymbol{\Omega}|\boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}) d\boldsymbol{\Omega} \\ &\propto (B_{km}^{-n} + \beta_m) \times \end{aligned}$$

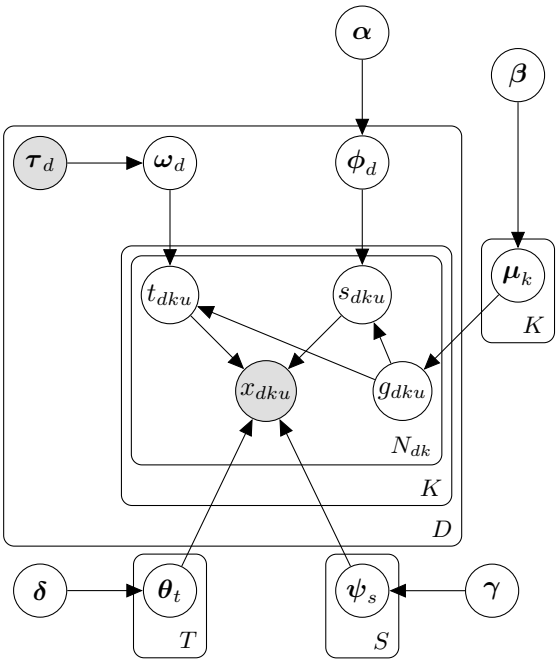

Figure 1: Plate notation of the model proposed in this paper (Eq. 2); see Fig. 2 for the notation.

- D Number of documents
- K Number of feature types (N_{dk}: of feature type k in document d)
- T Number of time bins
- S Number of background distributions
- $\boldsymbol{\theta}, \boldsymbol{\psi}$ Time-feature and background-feature proportions
- $\boldsymbol{\omega}, \boldsymbol{\phi}$ Document-time and document-background proportions
- $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\tau}$ Dirichlet priors
- $n := dku$ (document d, feature type k, occurrence u)
- Counters for the Gibbs Sampler:
  A_{ds} # genre s assigned to document d
  B_{km} # feature k generated by the time ($m = 0$) or the topic ($m = 1$) distributions.
  C_{sk} # feature k generated by genre s
  D_{tk} # feature k generated by time t
  E_{dt} # time t assigned to document d

Figure 2: Notation used in Fig. 1 and Eq. 2

$$\begin{cases} \frac{E_{dt}^{-n}+\tau_{dt}}{\sum_u^T E_{du}+\tau_{du}} \cdot \frac{D_{tk}^{-n}+\delta_k}{\sum_u^K D_{tu}^{-n}+\delta_u} & \text{if } g_n = 0 \\ \frac{A_{ds}^{-n}+\alpha_s}{\sum_u^S A_{du}^{-n}+\alpha_u} \cdot \frac{C_{sk}^{-n}+\gamma_k}{\sum_u^K C_{su}^{-n}+\gamma_u} & \text{else} \end{cases} \quad (2)$$

Since mixture models are often sensitive to the choice of hyperparameters (Wallach et al., 2009; Asuncion et al., 2009), $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}$ are updated after each iteration of the sampler using the estimates described by Minka (2003).

4. Data and features

4.1. Linguistic features

The data are extracted from the Digital Corpus of Sanskrit (DCS, Hellwig (2010 2020)), which contains more than 200 Sandhi-split texts in Vedic and Classical Sanskrit along with manually validated morphological and lexical information for each word.[1] The Vedic subcorpus of the DCS, as used in this paper, contains 35 texts with a total of 540,000 words. In contrast to previous philological work (see Sec. 2.), this paper uses a wide range of linguistic features (see Hellwig (2019, 4-7)), including, among others, the counts of the 1,000 most frequent words in the Vedic subcorpus of the DCS, cases, POS tags, verbal classes, tenses and moods.

As post-Rigvedic Sanskrit was not in active daily use, previous research has claimed that most linguistic changes took place in its vocabulary. Apart from the actual vocabulary, this paper therefore pays special attention to etymology[2] and derivational morphology, two word-atomic feature types that reflect changes on the lexical level. It has been claimed that post-Rigvedic Sanskrit incorporates an increasing amount of non-Indo-Aryan words due to its contact with substrate languages (Witzel, 1999), so that higher ratios of words with a non-Indo-Aryan etymology may indicate a later date of texts (Hellwig, 2010).

Derivational rules were used to derive new words (preferably nouns) from verbal stems and other nouns. Such processes can be as simple as using the verbal root as a noun or adjective (*diś-* 'to show' → *diś-* 'indication, direction'), but may also involve complex phonological transformations applied to already derived or compounded nouns (*sukara-* 'easy-to do' → *saukārya-* 'the state of being easy to do'). While Hellwig (2019) used only a limited amount of derivational information, this paper inspects the distribution of 84 rules based on the treatment in Wackernagel and Debrunner (1954). Lexicalizing compounds was another popular method for deriving new words; e.g. *saroruhāsana* = *saras-ruha-āsana* = 'lake-growing-seat' = 'having a lotus as his seat' = 'name of the god Brahman'. Previous research has not used the number of elements in such compounds systematically for studying the chronology of Sanskrit (a few brief notes in Wackernagel (1905, 6-9, 24-26)). Currently, etymological or derivational information is available for 61,5% of all Vedic word types. Derivational morphology and lexical compounding are mutually exclusive and are therefore subsumed under a single feature type "derivation".

Apart from these word-atomic features, two multi-word features are also considered. Recent research has provided evidence for an increasing degree of configurationality in Indo-Aryan, i.e. to use word order for marking grammatical functions (Reinöhl, 2016). As a syntactic treebank is only available for a small subset of Vedic texts (Hellwig et al., 2020), the most frequent 500 bi- and trigrams of POS tags are used as a coarse approximation of syntactic chunks (Hellwig, 2019). The second multi-word feature encodes the lengths of non-lexicalized compounds. While compounds in the ṚV and the AV have at most three members (Wackernagel, 1905, 25-26), their length is not limited in Classical Sanskrit (Lowe, 2015, 80-83), so that, as a working hypothesis, increasing counts of long compounds may be indicative of late Vedic texts.

Each text is split into sections of 200 words. Since each word contributes multiple atomic features (e.g. POS, derivational information) and forms part of POS bi- and tri-

[1] Conllu files are available from `https://github.com/OliverHellwig/sanskrit/tree/master/dcs/data/conllu`.

[2] This term is used here in its restricted meaning as "étymologie-origine"; see Mayrhofer (1992, IX-XIV).

grams, each text section contains 440 data points on average.

4.2. Temporal priors

The model described in Sec. 3. requires temporal priors $\boldsymbol{\tau}$ (see Eq. 1) that encode chronological proposals made in previous literature. Based on Renou (1957, 1-16), Witzel (1989), and Kümmel (2000, 5-6), this paper uses a fivefold temporal split of the VC:

Rigvedic (RV) 1300-1000 BCE; ṚV 1-9

Mantra language (MA) 1100-900 BCE; ṚV 10, Atharvaveda Saṃhitās, Ṛgveda-Khilāni, metrical parts of the Yajurveda Saṃhitās

Old prose (PO) 900-700 BCE; Aitareya Brāhmaṇa 1-5, Śatapatha Brāhmaṇa 6-9, 10.1-5; prose parts of the Yajurveda Saṃhitās

Late prose (PL) 700-400 BCE; major Brāhmaṇas not contained in PO, old Upaniṣads

Sūtra level (SU) 600-300 BCE; late Upaniṣads and Brāhmaṇas (e.g., the Gopatha Brāhmaṇa), the ritual handbooks called Sūtras

5. Evaluation

Section 5.1. studies the information that is encoded in the background distributions of ToB. Section 5.2. compares ToB with a baseline LDA model, using perplexity for the intrinsic and temporal predictions for the extrinsic evaluation. Here, the extrinsic evaluation is being complicated by the fact that the only diachronic information at our disposal is already encoded in the subjective priors $\boldsymbol{\tau}$. Section 5.3. takes a closer look at features that are generated by the time path of ToB, and discusses their philological relevance. The concluding Sec. 5.4. examines the temporal predictions for the ṚV.

5.1. The role of the background distributions

The background distributions are expected to capture the proportion of linguistic variation that cannot be explained by diachronic changes. In order to determine the optimal number of these background distributions, perplexity and accuracy are measured on the held out sets of cross-validations for varying numbers of background distributions.[3] As discussed in Hellwig (2019), randomly assigning text sections to the train and test sets underestimates the error rates on the test set of a discriminative model, because the linguistic evidence from the train sections is often strong enough to cause overfitting. Therefore the same splitting scheme as proposed in Hellwig (2019) ("textwise CV") is used in this paper. Here, each text is in turn used as the test set, and the model is trained with the remaining

[3]Accuracy is a short-hand term for the probability that the model prediction has been generated by the normal distribution that is derived from the coarse Vedic chronology given in Sec. 4.2.; see the discussion of $\boldsymbol{\tau}$ on p. 2. When the training is completed, section-wise date predictions for the left out text are obtained using "folding in".

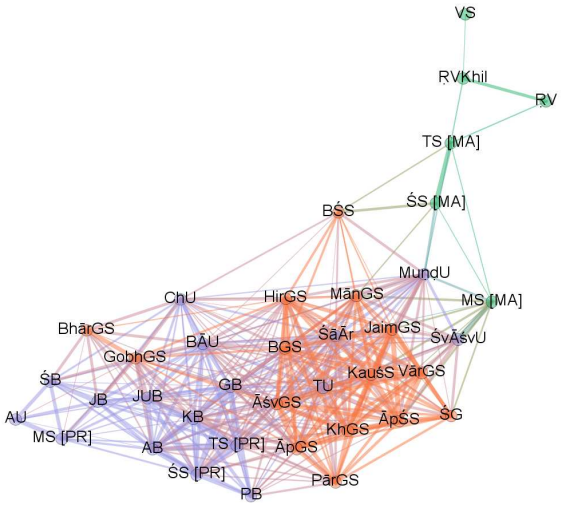

Figure 3: Undirected graph resulting from textwise similarities of background distributions; edge sizes are proportional to the textwise similarities. The graph induces a distinction between old metrical texts (top, green; ṚV, Ṛgveda-Khilāni, ŚS), prose texts (bottom left, blue; names ending on B = Brāhmaṇas) and ritual handbooks (bottom right, red; names ending on S[ūtra]). The Upaniṣads (names ending on U and Up) mediate between prose and Sūtras.

$T-1$ texts. When varying the number S of background distributions between 1 and 30, the setting $S=3$ results in the lowest perplexity and highest accuracy. This setting is used for all following experiments.

In order to understand which type of linguistic variation is encoded in the background, the counts of background assignments per text are accumulated and normalized, resulting in T distributions $\mathbf{b}_t$. The Euclidean distance between $\mathbf{b}_i$ and $\mathbf{b}_j$ is chosen for calculating the distance between a pair of texts (i,j). Using these Euclidean distances as edge weights results in the undirected graph that is shown in Fig. 3. The structure of the graph indicates a threefold split of the VC into early metrical texts (ṚV, Ṛgveda-Khilāni, ŚS), the works composed in prose and the ritual handbooks composed in the elliptic Sūtra style, which differs significantly from the style of other prose texts (Gonda, 1977, 629-647). Major Upaniṣads (esp. the Chāndogya Up. [ChU] and the Bṛhadāraṇyaka Up. [BĀU]) occupy an intermediate position between prose texts and Sūtras, although they were originally part of Brāhmaṇa texts. The structure of the graph therefore suggests that the background distributions primarily encode stylistic and genre-specific linguistic variation, as the differences in content between the three main groups go along with obvious differences in style.

5.2. Model comparison

While the evaluation of the background distributions (Sec. 5.1.) suggests that the text genre is a relevant factor when studying linguistic variation in Vedic, it cannot be taken for granted that ToB, the model proposed in this paper, is best suited for detecting time-dependent linguistic varia-

tion. ToB is therefore compared with a modified version of LDA (Blei et al., 2003) in which the flat prior of standard LDA is replaced with the subjective temporal prior τ of ToB.

For the intrinsic comparison, I perform textwise CVs (see Sec. 5.1.), using an uninformative temporal prior for each tested text, and compare the perplexities of the two models on the test texts using a pairwise Wilcoxon rank sum test. Under the alternative hypothesis that ToB has a lower perplexity than the baseline LDA, the test yields a highly significant p-value of $3.62e^{-8}$. The lower perplexity (i.e. higher likelihood) of ToB can be due to overfitting, as it has more parameters than LDA. Therefore the Bayesian Information Criterion (BIC; Schwarz (1978)), which penalizes higher numbers of parameters and thus favors plain LDA, is calculated for all tests. In around 70% of all cases, LDA has a higher BIC than ToB and is thus more appropriate than ToB according to this metric. Repeating the Wilcoxon test with the BIC values, however, yields a p-value of 0.016, which is not significant at the 1% level. When plotting the BIC values of LDA against those of ToB (not shown in this paper), it can be observed that for lower BICs ToB performs better than LDA. The respective texts are, in general, the earlier ones (ṚV, ŚS), and they contain samples of the Brāhmaṇa style, which may be more prone to textual interpolation than the Sūtra texts for which LDA has a lower BIC than ToB. A follow-up study should evaluate if this apparent correlation between time, genre and the BIC is systematic.

For performing an extrinsic comparison, it is evaluated how well the temporal range of each text (see Sec. 4. and Fn. 3) is predicted, again using uninformative temporal priors for each tested text. It is important to emphasize once more that these temporal ranges do not constitute a proper gold standard, because multiple historical strata can, in principle, occur in any text of the VC. A model that works correctly can therefore generate temporal predictions for individual sections of a text that massively deviate from the temporal priors. Keeping these restrictions in mind, the priors are again assumed to constitute Normal distributions (see Sec. 3.) and the z-standardized value of each prediction given the respective Normal distribution is calculated. In this scenario, values closer to 0 correspond to a better model fit. A Wilcoxon test that compares the z-values of both models (alternative hypothesis: ToB generates lower z-values than plain LDA) yields a p-value of less than $2.2e^{-16}$ and thus a highly significant result.

5.3. Time-correlated features

A central motivation for developing ToB is to extend the set of linguistic features that show systematic diachronic variation and can thus be used for dating and stratifying the VC (see Sec. 2.). The switch between temporal and background distributions in ToB (variable $\mathbf{g}$ in Fig. 1) can be used to find feature types that are predominantly generated by the time path of the model. When the feature types examined in this paper are ordered by the proportions with which they are generated by the time path of ToB, the top position is occupied by compounds (only generated by time), followed by infinite verbal forms (89,5%), lexical information (83,6%), tenses and modes (82,7%) and POS trigrams (76,5%). All remaining feature types are also preferably generated by the time path except for etymological information (39,2%).

The increasing use of compounds for expressing syntactic constructions including coordination, nominal subordination, and exocentric relations has often been described in secondary literature (Lowe, 2015). Since compounds with more than two components only appear in larger numbers at the end of the Vedic period (esp. in the Sūtra texts), this result is mainly relevant for dating texts composed in (early) Classical Sanskrit.

The important role of the lexicon and of finite verbal forms is not surprising, as these feature types have been used regularly in previous attempts to date early Vedic texts (e.g., Arnold (1905), Poucha (1942)). More interesting insights are provided by the POS n-grams. When plotting the POS type-token ratios (TTR) against the time slots predicted by the model (see Fig. 4), it can be observed that the TTRs of all POS n-grams are maximal for the ṚV and later on decrease with the predicted dates. This suggests that the syntactic variability of post-Rigvedic Sanskrit decreases as well, perhaps caused by processes of grammaticalization and configurationality which are in effect in Middle- and New Indo-Aryan languages (Heine and Reh (1984, 67), Reinöhl (2016)). It is also instructive to inspect the POS trigrams that are preferably associated with the two temporal extremes of the VC. In the earliest layer we find, for example, the sequence preverb – noun (in various cases) – finite verb (CADP-NC.*-V), which represents tmesis (i.e. separation of preverb and verb) in many passages such as the Soma hymn ṚV 9.86.31a (matching pattern underlined): *prá rebhá ety áti vā́ram avyáyam* "The husky-voiced one [= the Soma] goes forth across the sheep's fleece" (Jamison and Brereton, 2014, 1324); or, more frequently, with a noun in the accusative in central position (ṚV 10.67.12ab, about Indra's deeds): *índro mahnā́ maható arṇavásya ví mūrdhā́nam abhinad arbudásya* "Indra with his greatness split apart the head of the great flood, of Arbuda" (Jamison and Brereton, 2014, 1490). Eventual misassignments as at ṚV 9.73.2b (***ūrmā́v ádhi*** *venā́ avīvipan* 'the longing ones have made him (Soma) tremble on the wave'), where *ádhi-* 'on, in' is used as a postposition, but not as a preverb, could be avoided when a treebank of the complete VC is available. At the other end of the historical spectrum, late Vedic texts have a preference for absolutive constructions of compound verbs in clause final position (trigram NC.acc-CADP-CGDA), as at JUB 4.9.9: *prāṇebhyo 'dhi mṛtyupāśān unmucya-athainaṃ ... sarvamṛtyoḥ spṛṇāti* 'having released the fetters of death from his breaths, he releases him from all (kinds of) death'.

Temporal predictions for derivational features reflect many diachronic trends described in previous literature. When the derivational features are ordered by the mean date assigned to them, the first (= earliest) position is occupied by the suffix *-tāti*, which is used to derive abstract nouns from other nouns as in *sarvá-tāti-* 'complete-ness' (< *sárva-* 'complete, all') and known to be restricted to the oldest parts of the VC (Wackernagel and Debrunner, 1954,

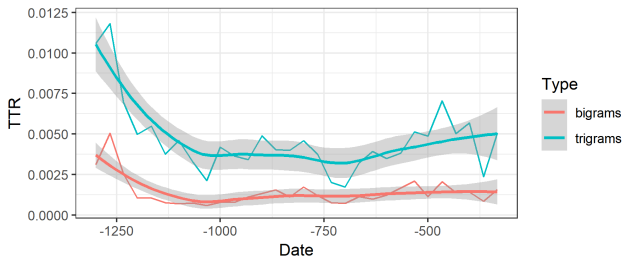

Figure 4: Type-token ratios of POS n-grams (y-axis) dependent from the predicted dates (x-axis). The curves demonstrate the decreasing syntactic variability of post-Rigvedic Sanskrit.

§464). Suffixes assigned to the latest time slots contain, among others, the comparative suffix *-tara* (e.g., *kṣipra-tara-* 'faster'), which replaces the older comparative suffix *-īyas* (e.g., *kṣépīyas-* 'faster', see Wackernagel and Debrunner (1954, §450)), or the suffix *-ika* with vṛddhi of the first syllable, which is often used to derive adjectives from (compounded) nouns (e.g., *aíkāh-ika-* 'lasting one day' < *eka-aha-* 'one day' with vṛddhi e → ai; see Wackernagel and Debrunner (1954, §194 b β) for a historical sketch). As often, results for the earliest Vedic strata are well known, while features associated with intermediate and late time ranges have the potential to promote philological research.

As mentioned on p. 3, lexicalized compounds are subsumed under the feature type derivation. Conforming to the general trend observed for compound formation (see above), the model assigns an earlier average date to words with two compound members (e.g., *vanas-pati-* 'lord of the wood, tree') than to those with three (e.g., *a-prajás-tā-* 'childlessness'). It should, however, be noted that 63% of the three-element compounds are inflected forms of the word *sv-iṣṭa-kṛt-* 'offering a good sacrifice', the name of a special sacrifice to the god Agni (Mylius, 1995, 140), which is almost exclusively discussed in the late Sūtra texts. Even this brief overview shows the importance of derivational information for inducing the temporal structure of the VC. Wüst (1928), who studied a related set of features for the ṚV, did not meet enthusiastic support in Vedic studies – it may be worthwhile to reconsider his approach with new quantitative methods.

5.4. Detail study: Temporal stratification of the Rigveda

The ṚV, the oldest work of Vedic Sanskrit, is a collection of ten books of religious poetry composed by multiple authors (Witzel, 1997, 261-264). Among all Vedic texts, the ṚV has been studied most intensively and can thus serve as a test case for the temporal predictions made by ToB. On the basis of linguistic criteria, citations, and the textual content, it is generally assumed that ṚV 10 is the youngest book of the whole collection (Renou, 1957, 4). The so-called Family Books (ṚV 2-7) are usually considered to be old or even to constitute the core of the ṚV (Witzel, 1997, 262-264). ṚV 9 is also often accepted as old, while the status of ṚV 1 and especially ṚV 8 is disputed (Hopkins (1896), Gonda (1975, 8-14), Jamison and Brereton (2014, 9-13)). Overall, the split (1-9) (10) has emerged as the most widely accepted

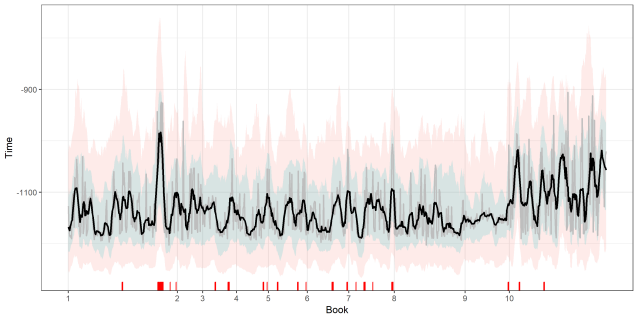

Figure 5: Predicted dates for the ṚV. The polygons show the smoothed 50% and 90% quantiles, the black line is the smoothed median, and the grey line is the unsmoothed median.

stratification of the ṚV.

Figure 5 shows the median and two quantiles of the dates predicted by ToB.[4] The overall trend observed in Fig. 5 confirms the most frequently postulated stratification of the ṚV: While book 10 is late, there are no clear temporal separations between the remaining nine books. For deriving a temporal ranking of the ten Rigvedic books, one-sided Wilcoxon rank sum tests between pairs of books (i, j) are performed. If the test for (i, j) is significant at the 10% level, an ordering constraint $i < j$ is recorded. When a minimum location shift of one time step is assumed, the resulting constraints induce the "canonical" ordering $(1 - 9) < (10)$. Leaving the location shift unspecified[5] induces the ordering $(4, 8) < (1 - 3, 5 - 7, 9) < (10)$, which deviates from the most widely accepted split (1-9) (10) by labeling RV 4 and 8 as the earliest books, as already postulated for book 8 by Lanman (1872, 580) and Arnold (1897a, 319) (strongly contested by Hopkins (1896)) and for book 4 by Wüst (1928).

Further binomial tests are performed for all features that are preferably assigned to the earliest time slots, assessing if they are significantly more frequent in RV 4 and 8 than in the rest of the text (RV 10 can be omitted as obviously younger). These tests produce a list of 92 features, most of which have been considered as archaic in previous research: (1) perfect subjunctive and injunctive (see Arnold (1905, 31)); (2) the suffixes *-tave*, *-vane*, *-aye* and *-ase*, all of which form dative verbal nouns (Wackernagel and Debrunner, 1954, s.v.); (3) the derivational suffixes *-tvana* (abstracts) and *-vat* (in *pra-vat-* 'elevation'; see Wackernagel and Debrunner (1954, §530,703)); (4) five POS n-grams containing, among others, the sequence noun-infinitive (as in old constructions like *jyók ca sū́ryaṃ dṛśé* 'in order to see the sun for a long time'); (5) and a list of 79 words.

In 1888, the scholar H. Oldenberg claimed that the hymns in each book of the ṚV are arranged according to the numbers of their stanzas, and that hymns violating this rule represent the youngest layer of Rigvedic poetry ("appendices"; Oldenberg (1888, 191-197, 265)). As Oldenberg's work is still among the most frequently cited studies on the textual

[4] Continuous quantiles are calculated by interpolating the discrete counts.

[5] Note that significant p-values can result from the mere sample sizes in this setting.

history of the ṚV, it may be useful to compare his results with the output of ToB. The 31 hymns identified as appendices in Oldenberg (1888, 197-202, 222-223) are marked by the rug plot at the bottom of Fig. 5, and obviously coincide with some of the peaks in the predicted times.[6] A Wilcoxon rank sum test that compares the times predicted for Oldenberg's appendices with those of the rest of ṚV 1-7, 9 produces a highly significant p-value of less than $2e1^{-16}$, which suggests that Oldenberg's ideas are supported by the output of ToB. A closer inspection, however, shows that this strong effect is mainly caused by a few of Oldenberg's appendices marked as especially young by the model. These hymns comprise, among others, ṚV 1.162-164 (including the famous "riddle hymn" 1.164, which may be related to the pravargya ritual; see Houben (2000)); the "frog hymn" 7.103, which shows traits of later religious ideas (Lubin, 2001); the Atharvanic hymn ṚV 7.104 (Lommel, 1965, 203ff.); the Soma hymn 9.113, which foreshadows a concept of heaven occurring in much later texts (Jamison and Brereton, 2014, 1304) and notably mentions a group of Gandharvas instead of a single Gandharva only, an idea often considered as late (Oberlies, 2005, 106); 10.19, a hymn composed in easy language that addresses cows who have gone astray, but is found, somehow unfittingly, at the end of a series of funeral hymns (Jamison and Brereton, 2014, 1401); and 10.60, which pays much attention to the Atharvanic topic of healing. The remaining appendices, esp. those contained in the Family Books ṚV 2-7, are not marked as particularly late by the model, but some of them even as quite old as, for example, the "praise of giving" (*dānastuti-*) in ṚV 5.27, whose status as an appendix has been challenged by Jamison and Brereton (2014, 688) on metrical grounds.

6. Summary

This paper has introduced a Bayesian mixture model with a temporal component that is used for chronological research in Vedic literature. Although the VC is used as the text corpus in this paper, the proposed method is not specifically designed for Vedic Sanskrit, but can be applied to any corpus with a disputed historical structure as long as linguistic annotations for this corpus are available. As Sections 3. and 5. have shown, the actual challenge is rather the evaluation of such a model than its design. While the underlying probabilistic processes are well understood, the interpretation of the model output requires a close interaction between quantitative methods and text-historical scholarship, especially since the data with which the model are evaluated do not constitute a proper gold standard (see Sec. 5.1. and 5.2.). The brief evaluation of the ṚV in Sec. 5.4. functions as a test case that indicates some possible approaches. Although a closer inspection of the results for the ṚV will unveil more insights into its structure, more interesting candidates for in-depth studies are certainly found among the post-Rigvedic texts as, for example, the two recensions of the Atharvaveda (see Whitney and Lanman (1905, cxxvii-xclii) and Witzel (1997, 275-284)) or early prose treatises such as the Maitrāyaṇī-Saṃhitā (see Amano (2009, 1-6) on the state of research).

On the mathematical side, the model proposed in this paper is a prototype that can be extended in various aspects. Its most serious drawback is the inflexible structure of the admixture models, which will be replaced by a Hierarchical Dirichlet Process (HDP, Teh et al. (2005)) in a follow-up study. In addition, the fixed size of the text windows (see Sec. 4.1.) prevents textual strata from being directly induced from the data (instead of constructing them in a post-processing step). Combining HDPs with a Markov Random Field, as proposed by Orbanz and Buhmann (2008) for image segmentation, appears to provide a viable solution for this challenge.

[6] Only full hymns marked as appendices are considered in this paper, i.e. ṚV 1.104, 162-164, 179, 191; 2.42-43; 3.28-29, 52-53; 4.48, 58; 5.27-28, 61, 87; 6.47, 74-75; 7.17, 33, 55, 103-104; 9.112-114; 10.19, 60.

7. Bibliographical References

Amano, K. (2009). *Maitrāyaṇī Saṃhitā I–II. Übersetzung der Prosapartien mit Kommentar zur Lexik und Syntax der älteren vedischen Prosa*. Hempen, Bremen.

Anand, D. A. and Jana, S. (2013). Chronology of Sanskrit texts: An information-theoretic corroboration. In *National Conference on Communications (NCC)*, pages 1–5. IEEE.

Arnold, E. V. (1897a). Literary epochs in the Rigveda. *Zeitschrift für vergleichende Sprachforschung auf dem Gebiete der Indogermanischen Sprachen*, 34(3):297–344.

Arnold, E. V. (1897b). Sketch of the historical grammar of the Rig and Atharva Vedas. *Journal of the American Oriental Society*, 18:203–353.

Arnold, E. V. (1905). *Vedic Metre in its Historical Development*. University Press, Cambridge.

Asuncion, A., Welling, M., Smyth, P., and Teh, Y. W. (2009). On smoothing and inference for topic models. In *Proceedings of the Twenty-fifth Conference on Uncertainty in Artificial Intelligence*, pages 27–34. AUAI Press.

Avery, J. (1872). Contributions to the history of verb-inflection in Sanskrit. *Journal of the American Oriental Society*, 10:219–324.

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 113–120.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Chemudugunta, C., Smyth, P., and Steyvers, M. (2007). Modeling general and specific aspects of documents with a probabilistic topic model. In *Advances in Neural Information Processing Systems*, pages 241–248.

Falk, H. (1993). *Schrift im alten Indien: Ein Forschungsbericht mit Anmerkungen*. Gunter Narr Verlag, Tübingen.

Fosse, L. M. (1997). *The Crux of Chronology in Sanskrit Literature*. Scandinavian University Press, Oslo.

Frermann, L. and Lapata, M. (2016). A Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Gill, P. S. and Swartz, T. B. (2011). Stylometric analyses using Dirichlet process mixture models. *Journal of Statistical Planning and Inference*, 141(11):3665–3674.

Gonda, J. (1975). *Vedic Literature (Saṃhitās and Brāhmaṇas)*, volume 1 of *A History of Indian Literature*. Otto Harrassowitz, Wiesbaden.

Gonda, J. (1977). *The Ritual Sūtras*, volume 1, Fasc. 2 of *A History of Indian Literature*. Otto Harrassowitz, Wiesbaden.

Hall, D., Jurafsky, D., and Manning, C. D. (2008). Studying the history of ideas using topic models. In *Proceedings of the EMNLP*, pages 363–371.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 2116. NIH Public Access.

Heine, B. and Reh, M. (1984). *Grammaticalization and Reanalysis in African Languages*. Helmut Buske Verlag, Hamburg.

Hellwig, O., Scarlata, S., Ackermann, E., and Widmer, P. (2020). The treebank of Vedic Sanskrit. In *Proceedings of the LREC*.

Hellwig, O. (2010). Etymological trends in the Sanskrit vocabulary. *Literary and Linguistic Computing*, 25(1):105–118.

Hellwig, O. (2010–2020). DCS - The Digital Corpus of Sanskrit. `http://www.sanskrit-linguistics.org/dcs/index.php`.

Hellwig, O. (2019). Dating Sanskrit texts using linguistic features and neural networks. *Indogermanische Forschungen*, 124:1–47.

Hock, H. H. (2000). Genre, discourse, and syntax in early Indo-European, with emphasis on Sanskrit. In Susan C. Herring, et al., editors, *Textual Parameters in Older Languages*, pages 163–196. John Benjamins, Amsterdam/Philadelphia.

Hoffmann, K. (1967). *Der Injunktiv im Veda*. Winter, Heidelberg.

Hopkins, E. W. (1896). Prāgāthikāni, I. *Journal of the American Oriental Society*, 17:23–92.

Houben, J. E. M. (2000). The ritual pragmatics of a Vedic hymn: The "Riddle Hymn" and the pravargya ritual. *Journal of the American Oriental Society*, 120(4):499–536.

Jamison, S. W. and Brereton, J. P. t. (2014). *The Rigveda: the Earliest Religious Poetry of India*. Oxford University Press, New York.

Jamison, S. W. (1991). Syntax of direct speech in Vedic. pages 40–56. E.J. Brill, New York, Kopenhagen, Köln.

Kim, Y., Chiu, Y.-I., Hanaki, K., Hegde, D., and Petrov, S. (2014). Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65.

Kümmel, M. J. (2000). *Das Perfekt im Indoiranischen. Eine Untersuchung der Form und Funktion einer ererbten Kategorie des Verbums und ihrer Entwicklung in den altindoiranischen Sprachen*. Reichert, Wiesbaden.

Lanman, C. R. (1872). A statistical account of noun-inflection in the Veda. *Journal of the American Oriental Society*, 10:325–601.

Levitt, S. H. (2003). The dating of the Indian tradition. *Anthropos*, 98(2):341–359.

Lommel, H. (1965). Vasiṣṭha und Viśvāmitra. *Oriens*, 18/19:200–227.

Lowe, J. J. (2015). The syntax of Sanskrit compounds. *Language*, 91(3):71–115.

Lubin, T. (2001). Vratá divine and human in the early Veda. *Journal of the American Oriental Society*, 121(4):565–579.

Masica, C. P. (1991). *The Indo-Aryan Languages*. Cambridge University Press, Cambridge.

Mayrhofer, M. (1992). *Etymologisches Wörterbuch des Altindoarischen. I. Band*. Carl Winter, Heidelberg.

Mikros, G. and Argiri, E. (2007). Investigating topic influence in authorship attribution. In *Proceedings of the International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection*, pages 29–35.

Minka, T. P. (2003). Estimating a Dirichlet distribution. Technical report.

Mylius, K. (1995). *Wörterbuch des altindischen Rituals*. Institut für Indologie, Wichtrach.

Oberlies, T. (2005). Der Gandharva und die drei Tage währende 'Quarantäne'. *Indo-Iranian Journal*, 48(1):97–109.

Oldenberg, H. (1888). *Die Hymnen des Ṛigveda. Band I: Metrische und textgeschichtliche Prolegomena*. Wilhelm Hertz, Berlin.

Orbanz, P. and Buhmann, J. M. (2008). Nonparametric Bayesian image segmentation. *International Journal of Computer Vision*, 77(1-3):25–45.

Poucha, P. (1942). Schichtung des Ṛigveda. *Archív Orientální*, 13:103–141, 225–269.

Rau, W. (1983). *Zur vedischen Altertumskunde*. Steiner, Wiesbaden.

Reinöhl, U. (2016). *Grammaticalization and the Rise of Configurationality in Indo-Aryan*. Oxford University Press, Oxford, UK.

Renou, L. (1957). *Altindische Grammatik, Introduction Générale*. Vandenhoeck & Ruprecht, Göttingen.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 487–494.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Seroussi, Y., Bohnert, F., and Zukerman, I. (2012). Authorship attribution with author-aware topic models. In *Proceedings of the ACL: Short Papers-Volume 2*, pages 264–269. Association for Computational Linguistics.

Stamatatos, E. (2009). A survey of modern authorship at-

tribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2005). Sharing clusters among related groups: Hierarchical Dirichlet processes. In *Advances in Neural Information Processing Systems*, pages 1385–1392.

Wackernagel, J. and Debrunner, A. (1954). *Altindische Grammatik. II, 2: Die Nominalsuffixes*. Vandenhoeck & Ruprecht, Göttingen.

Wackernagel, J. (1905). *Altindische Grammatik. Band II, 1: Einleitung zur Wortlehre. Nominalkomposition*. Vandenhoeck & Ruprecht, Göttingen.

Wallach, H. M., Mimno, D. M., and McCallum, A. (2009). Rethinking LDA: Why priors matter. In *Advances in Neural Information Processing Systems*, pages 1973–1981.

Wang, X. and McCallum, A. (2006). Topics over time: A non-Markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International conference on Knowledge Discovery and Data Mining*, pages 424–433.

Whitney, W. D. and Lanman, C. R. (1905). *Atharva-Veda Samhita*. Harvard University, Cambridge.

Witzel, M. (1989). Tracing the Vedic dialects. In Colette Caillat, editor, *Dialectes dans les littératures indo-aryennes*, pages 97–264. Collége de France, Institut de Civilisation Indienne, Paris.

Witzel, M. (1995). Early Indian history: Linguistic and textual parametres. In George Erdosy, editor, *The Indo-Aryans of Ancient South Asia. Language, Material Culture and Ethnicity*, volume 1, pages 85–125. Walter de Gruyter, Berlin, New York.

Witzel, M. (1997). The development of the Vedic canon and its schools: The social and political milieu (Materials on Vedic Sakhas, 8). In Michael Witzel, editor, *Inside the Texts, Beyond the Texts. New Approaches to the Study of the Vedas*, pages 258–348. Cambridge.

Witzel, M. (1999). Substrate languages in Old Indo-Aryan (Ṛgvedic, Middle and Late Vedic). *Electronic Journal of Vedic Studies*, 5(1):1–67.

Wüst, W. (1928). *Stilgeschichte und Chronologie des Ṛgveda*. Deutsche Morgenländische Gesellschaft, Leipzig.